Psychologische und sozialwissenschaftliche Kurzskalen für Wissenschaft und Praxis – Eine Einführung

Christoph J. Kemper, Elmar Brähler und Markus Zenger

Kurzskalen für die Messung menschlichen Erlebens und Verhaltens

Das Interesse, menschliches Erleben und Verhalten mit zeitökonomischen Erhebungsverfahren[1] zu messen, ist fast ebenso alt, wie die ersten Testverfahren in den Anfängen der psychometrischen Persönlichkeitsdiagnostik vor mehr als 100 Jahren. Bereits wenige Jahre nach der Einführung des ersten psychometrischen Tests, des Binet-Simon-Tests zur Messung der Intelligenz von Schulkindern (Binet, 1905), schlug Doll (1917) vor, dass für die Messung von Intelligenz auch weniger Testaufgaben (Items) ausreichen, als im ursprünglichen Binet-Simon-Test vorgesehen. Seitdem scheint das Interesse an zeitökonomischen Testverfahren (Kurzskalen[2]) ungebrochen. Dieser Trend war zunächst auf die Messung kognitiver Persönlichkeitsmerkmale beschränkt, breitete sich aber zunehmend auch auf die Messung nicht-kognitiver Persönlichkeitsmerkmale aus (Coste, Guillemin, Pouchot & Fermanian, 1997; Rammstedt & John, 2007; Thalmayer, Saucier & Eigenhuis, 2011).

Angesichts der zahlreichen Vorteile, die Kurzskalen zur Messung von Persönlichkeit[3] bieten, ist deren Popularität verständlich. Für viele Anwender von Kurzskalen stehen Zeit- und Kostenersparnis im Vordergrund, die gegenüber umfangreichen Erhebungsverfahren erzielt werden können. Viele psychologische Testverfahren wurden spezifisch für den Einsatz in der psychologischen Einzelfalldiagnostik entwickelt. Aufgrund ihres Umfangs sind diese für bestimmte Anwendungskontexte der Persönlichkeitsmessung daher wenig geeignet (Rammstedt, Kemper & Schupp, 2013). Zum Beispiel ist psychologische, sozialwissenschaftliche und wirtschaftswissenschaftliche Forschung heute überwiegend multivariater Natur. Hinzu kommt, dass empirische Erhebungen oft interdisziplinär angelegt sind. Für die Messung einzelner Konstrukte steht in diesen Erhebungskontexten relativ wenig Erhebungszeit zur Verfügung, weil viele verschiedene Konstrukte gemessen werden müssen. Ein ähnliches Problem ergibt sich z. B. auch in der klinisch-psychologischen Einzelfalldiagnostik. Um die vielfältigen klinisch-relevanten Facetten der Psyche eines Patienten abbilden zu können, müssen oft viele unterschiedliche Tests vorgegeben werden (Stieglitz, 2007). Ziel dabei ist, sich zunächst schnell, effizient und ohne größere Belastung des Getesteten einen überblicksartigen Gesamteindruck von dessen Problembereichen machen zu können, die anschließend gezielt mit umfassenderen Testverfahren untersucht werden können. In den genannten Anwendungskontexten wird durch den Einsatz von Kurzskalen bspw. vermieden, dass eine Erfassung der Persönlichkeit aus Zeit- oder Kostengründen gänzlich unterbleibt oder negative Effekte von Persönlichkeitsmessungen, wie zufälliges Antworten, ungenaue Angaben, geringe Teilnahmebereitschaft oder hohe Abbrecherquoten, auftreten (Burisch, 1984; Credé, Harms, Niehorster & Gaye-Valentine, 2012; Edwards, Roberts, Sandercock & Frost, 2004; Robins, Hendin & Trzesniewski, 2001).

Der Einsatz von Kurzskalen zur Messung von Persönlichkeit findet allerdings nicht nur Fürsprecher. Die Nutzung von Kurzskalen ist nicht kompatibel mit einer seit Jahrzehnten etablierten Lehrmeinung, nach der ein reliabler und valider Test viele Items enthalten müsse (z. B. Nunnally, 1978), von denen jedes einzelne zur Reliabilität des Testwerts beiträgt und einen relevanten Teilaspekt des Konstrukts abbildet (vgl. z. B. Kruyen, Emons & Sijtsma, 2013). Gemäß dieser Position führe die Kürzung eines psychometrisch fundierten Tests unweigerlich zu einer verminderten Testgüte, die nicht vertretbar sei. Die ungünstigen psychometrischen Implikationen der Nutzung von Kurzskalen werden von Credé et al. (2012) näher ausgeführt: Eine kurze Skala enthält im Vergleich zu einer langen Skala mit gleichem Gültigkeitsanspruch

1 Erhebungsverfahren, Erhebungsinstrument, Testverfahren, diagnostisches Verfahren, Skala, Ratingskala und Fragebogen werden im Folgenden als Synonyme für „psychologischer Test“ (für eine Definition siehe S. 4) verwendet.

2 Für eine Definition von „Kurzskala“ siehe S. 4.

3 Für eine Definition von „Persönlickeit“ siehe S. 3.

weniger Items. Zufällige Messfehler der Items können sich weniger gut ausgleichen. Die Reliabilität des Testkennwerts ist daher niedriger und somit auch dessen Kriteriumsvalidität. Darüber hinaus wird die Kriteriumsvalidität ebenfalls durch eine geringere Konstruktabdeckung gemindert, da weniger Teilaspekte des Konstrukts durch die Messung abgebildet werden. Daher sinken Inhaltsvalidität und Kriteriumsvalidität der Kurzskala. Credé et al. (2012) führen weiter aus, dass der Einsatz von Kurzskalen die inferenzstatistische Überprüfung von Hypothesen negativ beeinflussen könnte, da die verminderte psychometrische Güte mit einer erhöhten Wahrscheinlichkeit für Fehler erster und zweiter Art einhergehen kann. Ein weiterer Einwand, der wiederholt vorgebracht wird, richtet sich nicht gegen Kurzskalen per se, sondern gegen deren Konstruktions- und Validierungsprozess (vgl. Smith, McCarthy & Anderson, 2000). Diese erfüllen häufig methodische Minimalstandards nicht. Beispielsweise werden fragwürdige Kürzungsstrategien, wie die Maximierung der internen Konsistenz, eingesetzt, suboptimale Validierungsstrategien angewendet, Validitätsbelege von der Langversion einer Skala auf eine Kurzversion übertragen oder schlicht gar keine Validitätsbelege für eine Kurzskala angegeben (Coste et al., 1997; Kruyon et al., 2013; Levy, 1968, Smith et al., 2000). Oft fehlt in den Veröffentlichungen auch eine kritische Reflektion von Stärken und Schwächen einer konstruierten Kurzskala (Kruyon et al., 2013).

Angesichts dieser Einwände erscheint die Forderung, die psychometrische Güte eines Tests nicht auf dem Altar der Effizienz zu opfern, allzu verständlich. Allerdings lassen sich auch empirische Studien anführen, welche die Verluste an psychometrischer Güte bei der ökonomischen Operationalisierung mancher Konstrukte als vertretbar erscheinen lassen. In einer aktuellen Studie verglichen Thalmayer et al. (2011) lange und gekürzte Fassungen gängiger Persönlichkeitstests (Big Five Inventory, International Personality Item Pool, HEXACO Personality Inventory) im Hinblick auf ihre Prädiktionskraft. Diese Autoren finden lediglich marginale Unterschiede bei der Vorhersage von Leistungs- und diversen Verhaltensmaßen innerhalb der jeweiligen Testfamilie und konstatieren: „Holding questionnaire and number of scales constant, and examining scores on briefer vs. longer versions for relative predictive ability, our results provide no support for using longer versions … Some important variation in psychological differences can be captured by self-report inventories, but it may be of a finite amount, sufficiently obtainable by a few high-validity items“ (p. 1006). Neben dieser Studie lassen sich weitere anführen, die zeigen, dass manche Kurzskalen trotz ihres geringen Umfangs eine akzeptable Retest-Reliabilität und Kriteriumsvalidität aufweisen (Burisch, 1984; Gosling, Rentfrom & Swann, 2003; Kemper, Lutz & Neuser, 2011; Paunonen & Jackson, 1985; Robins et al., 2001; Wood, Nye, & Saucier, 2010).

Die Frage, die angesichts der oben angeführten konzeptuellen Überlegungen und empirischen Befunde wünschenswerterweise gestellt werden sollte, ist daher nicht „ob“ die Verwendung von Kurzskalen prinzipiell legitim ist, sondern „unter welchen Bedingungen“ diese eingesetzt werden können; genauer gesagt, „welche Voraussetzungen“ vor dem Hintergrund eines bestimmten Anwendungskontextes gegeben sein sollten. Diese Entscheidung kann und muss von Fall zu Fall – also für jede Operationalisierung – auf der Grundlage der empirischen Befunde zur psychometrischen Güte der Kurzskala erfolgen. Damit die psychometrische Güte von Kurzskalen möglichst hoch ist und deren Anwendungsbereich möglichst wenige Einschränkungen erfährt, wäre es ebenfalls wünschenswert, wenn sich Entwickler von Kurzskalen an gängigen Standards und Richtlinien für die Entwicklung und Validierung von Testverfahren im Allgemeinen bzw. von Kurzskalen im Besonderen orientieren würden (z. B. Testbeurteilungssystem des Diagnostik- und Testkuratoriums der Föderation Deutscher Psychologenvereinigungen; Empfehlungen des Rats für Sozial- und Wirtschaftsdaten; siehe auch Credé et al., 2012; Marsh, Ellis, Parada, Richards & Heubeck, 2005; Stanton, Sinar, Balzer & Smith, 2002; Smith et al., 2000).

Zielsetzung und Zielgruppe des Buches

Das vorliegende Buch stellt – erstmals im deutschsprachigen Raum – potenziellen Anwendern eine umfassende Sammlung von zeitökonomischen und standardisierten Erhebungsverfahren für die Messung von menschlichem Erleben und Verhalten, u. a. Persönlichkeit, Leistung, Einstellungen und Psychopathologie, zur Verfügung. Die Erhebungsverfahren werden in kurzen, einheitlich und übersichtlich struk-

turierten Beiträgen mit Informationen zu Entwicklung, Aufbau, Anwendung, psychometrischer Güte und Verfügbarkeit vorgestellt. Dies erlaubt dem Leser eine schnelle Orientierung und Auswahl eines für seine Zwecke angemessenen Verfahrens (siehe Hinweise zum Gebrauch des Buches). Die vorgestellten Verfahren wurden mehrheitlich für die psychologische oder sozialwissenschaftliche Forschung entwickelt und in entsprechenden Fachzeitschriften veröffentlicht. Die Nutzung für nicht-kommerzielle Zwecke ist daher in der Regel kostenfrei. Ein weiteres Ziel des Buches ist es, den Entwicklungsbedarf im Hinblick auf die Operationalisierung von Konstrukten, die bisher nicht mit Kurzskalen erfasst werden, aufzuzeigen und auf diese Weise die Konstruktion von ökonomischen Erhebungsverfahren zu stimulieren. Schließlich soll dieses Buch auch zur Qualitätssicherung von psychologischen und sozialwissenschaftlichen Erhebungsverfahren beitragen. Die vorgestellten Testverfahren müssen gemäß den Selektionskriterien (siehe unten) an mindestens einer empirischen Stichprobe validiert worden sein. Ein Mindestmaß an psychometrischer Güte ist demnach sichergestellt. Darüber hinaus fördert eine weite Verbreitung und Anwendung dieser standardisierten Testverfahren, im Gegensatz zu ad-hoc-konstruierten studienspezifischen Verfahren, die Vergleichbarkeit von empirischen Untersuchungen und Befunden (vgl. Kemper, Beierlein, Bensch, Kovaleva & Rammstedt, 2012).

Das vorliegende Buch richtet sich an alle Personen, die aus verschiedensten Gründen ökonomische und standardisierte Erhebungsverfahren für die Erfassung menschlichen Erlebens und Verhaltens benötigen, z. B. Dozenten, Studierende, Berater, Surveymanager, Marktforscher, insbesondere aber an Wissenschaftler aus unterschiedlichen Disziplinen und Berufsgruppen. Dazu gehören Psychologen, Pädagogen und Mediziner, Gesundheitswissenschaftler, Sozialwissenschaftler (z. B. Umfrageforscher), Wirtschaftswissenschaftler (z. B. Verhaltensökonomen), Politikwissenschaftler usw. Außerdem sind viele der im Buch besprochenen Kurzskalen prinzipiell als Screeninginstrumente für diverse gesundheitsbezogen arbeitende Berufsgruppen interessant, z. B. für Psychotherapeuten, Pädagogen und Mediziner verschiedener Facharztrichtungen.

Selektionskriterien

Inhaltsbereich

Zu den wichtigsten Informationsquellen für die Erklärung, Beschreibung und Vorhersage des Erlebens und Verhaltens von Menschen und dessen Auswirkung auf gesellschaftliche Prozesse und Phänomene gehören psychologische Merkmale, insbesondere Persönlichkeitsmerkmale (Roberts et al., 2007). Das vorliegende Buch thematisiert die Messung und Erfassung von Persönlichkeitsmerkmalen im weitesten Sinne. Um den unterschiedlichen Perspektiven auf menschliches Erleben und Verhalten, z. B. der psychologischen und der soziologischen Perspektive, Rechnung tragen zu können und zudem möglichst viele dafür relevante Erhebungsverfahren aufnehmen zu können, wurde eine sehr allgemeine Definition von Persönlichkeit als Einschlusskriterium zugrunde gelegt. Von Pervin, Cervone und John (2005, S. 31) wird Persönlichkeit definiert als „… jene Charakteristika oder Merkmale des Menschen, die konsistente Muster des Fühlens, Denkens und Verhaltens ausmachen“. Hierunter fallen bspw. Persönlichkeitsmerkmale im engeren Sinne wie Extraversion oder Neurotizismus, Fähigkeiten wie Intelligenz, Handlungsüberzeugungen wie Optimismus oder Selbstwirksamkeit, Einstellungen wie Ausländerfeindlichkeit oder Sexismus, klinische Merkmale wie Ängste, Depression oder Essstörungen und diverse weitere Merkmale und Merkmalsbereiche.

Gegenstände

Im vorliegenden Buch werden 95 psychologische und sozialwissenschaftliche Erhebungsverfahren vorgestellt, die eine ökonomische und effiziente Messung der Persönlichkeit anhand von Kurzskalen erlauben. Die Kurzskalen wurden bei Recherchen in einschlägigen Datenbanken wie Psyndex Tests oder Google

Scholar gefunden. Weiterhin wurden den Herausgebern bekannte Entwickler von Kurzskalen gezielt angesprochen. Um eine einheitliche Definition der Gegenstände zu ermöglichen, werden diese fortan als „psychologische Tests" bezeichnet: „Ein Test ist ein wissenschaftliches Routineverfahren zur Untersuchung eines oder mehrerer empirisch abgrenzbarer Persönlichkeitsmerkmale mit dem Ziel einer möglichst quantitativen Aussage über den relativen Grad der individuellen Merkmalsausprägung" (Lienert, 1969, S. 7). Hierzu gehören beispielsweise Persönlichkeitstests, Leistungstests, Einstellungstests und klinische Testverfahren (Screeningfragebögen, Beschwerdenlisten, Symptomfragebögen etc.). Als Synonyme für „psychologischer Test" werden die Begriffe Testverfahren, diagnostisches Verfahren, Erhebungsverfahren, Erhebungsinstrument, Skala, Ratingskala und Fragebogen verwendet. Nicht in das Buch aufgenommen wurden Indizes nach formativem Messkonzept, projektive Verfahren, Interviews, Klassifikationssysteme und apparative Verfahren.

Neben der Definition für psychologische Tests wurde der Begriff „Messung" für eine weitere Spezifikation und Eingrenzung der Gegenstände herangezogen. Unter einer Messung versteht Rost (2004), dass nicht-beobachtbare Sachverhalte – Konstrukte bzw. Persönlichkeitsmerkmale – über mindestens zwei manifeste Indikatoren erfasst werden müssen. Bei jedem im Buch vorgestellten Test sollten daher mindestens zwei Itemantworten nach psychometrisch geprüften Regeln zu einem mindestens ordinalen Test- bzw. Skalenwert aggregiert werden können (vgl. auch Barkmann, Schulte-Markwort & Brähler, 2011). Mit dieser Definition von Messung wurden die Tests auf eine Mindestanzahl von zwei Items pro Messung bzw. Test eingegrenzt.

Die Maximalanzahl an Items der hier vorgestellten Tests ergab sich aus der Beschäftigung mit dem Begriff Kurzskala bzw. dem Testgütekriterium der Ökonomie. Nach Bühner (2011) kann die Ökonomie eines Tests nur im Vergleich zu anderen Tests oder anderen diagnostischen Verfahren beurteilt werden. Eine Kurzskala kann demnach als ein psychologischer Test angesehen werden, der die ökonomische Messung eines Persönlichkeitsmerkmals ermöglicht, indem die Bearbeitungsdauer einer Testung im Vergleich zu längeren Tests mit ähnlichem Gültigkeitsanspruch reduziert wird. Ob eine Messung ökonomisch ist oder nicht, hängt demnach von anderen verfügbaren Operationalisierungen für ein Konstrukt ab; außerdem von dessen Breite und dem Kontext der Messung. Streng genommen müsste daher für jedes Konstrukt einzeln entschieden werden, ob eine bestimmte Messung ökonomisch ist oder nicht. Da dieser Ansatz wenig praktikabel wäre, wurde eine andere Strategie verfolgt. Um Richtwerte für die Maximalanzahl der Items zu erhalten, wurde ein für den Persönlichkeitsbereich etabliertes Testverfahren herangezogen, das NEO-Persönlichkeitsinventar in der revidierten Fassung (NEO-PI-R, Ostendorf & Angleitner, 2004). Das NEO-PI-R erfasst die im Fünf-Fakoren-Modell der Persönlichkeit (Norman, 1963; Tupes & Cristal, 1961; Costa & McCrae, 1985) postulierten abstrakten Persönlichkeitsmerkmale (auch: Domänen) Extraversion, Verträglichkeit, Gewissenhaftigkeit, Neurotizismus und Offenheit für Erfahrung. Außerdem weist jede Domäne eine Binnendifferenzierung auf. So gliedert sich bspw. die Domäne Extraversion im NEO-PI-R in mehrere Facetten: Herzlichkeit, Geselligkeit, Durchsetzungsfähigkeit, Aktivität, Erlebnishunger und Frohsinn. Jede Facette ist als eindimensional konzipiert und wird mit acht Items erfasst. Da die Facettenskalen des NEO-PI-R Persönlichkeitsmerkmale mit niedrigem bis mittleren Abstraktionsgrad erfassen, wurden als Selektionskriterium für die Tests des Buches acht Items als Obergrenze bei eindimensionalen Tests festgelegt. Zur Festsetzung der Maximalanzahl bei mehrdimensionalen Tests wurde ebenfalls ein Verfahren der NEO-Familie herangezogen. Das NEO-Fünf-Faktoren-Inventar (NEO-FFI, Borkenau & Ostendorf, 1993) ist eine gekürzte Fassung des NEO-PI-R. Im NEO-FFI werden im Gegensatz zum NEO-PI-R nur die Domänen, Persönlichkeitsmerkmale von hohem Abstraktionsgrad, mit jeweils 12 Items erhoben. Daher wurde die Maximalanzahl an Items für mehrdimensionale Tests auf 12 gesetzt.

Die obigen Eingrenzungen führten zum Ausschluss von Ein-Item-Skalen und eindimensionalen Tests mit mehr als acht bzw. mehrdimensionalen Tests mit mehr als 12 Items. Die Ausnahme bildeten Tests zur Erfassung der kognitiven Leistungsfähigkeit wie der BEFKI GC-K und der HMT-S, da zur Messung von kognitiven Persönlichkeitsmerkmalen wie Intelligenz üblicherweise deutlich mehr Items notwendig sind, als für die Erfassung anderer nicht-kognitiver Persönlichkeitsmerkmale. Ausnahmen im Hinblick auf die Obergrenze der Items wurden bei den Testverfahren VOCT, EVOS, NSA und NPI-10 gemacht. Der VOCT (siehe S. 340) wird zwar wie ein nicht-kognitiver Persönlichkeitstest ausgewertet, das Testmaterial ent-

spricht allerdings dem eines kognitiven Leistungstests. Das Testverfahren EVOS (siehe S. 72) enthält insgesamt 13 Items, allerdings gehen nur 12 in die zentralen Testkennwerte ein. NSA (siehe S. 212) erfasst einen Testkennwert anhand von fünf Szenarien. Aufgrund des Messkonzepts wurden die Szenarien als Items behandelt. Beim NPI-10 (siehe S. 204) wurden trotz der eindimensionalen Struktur mehr als acht Items akzeptiert, da Narzissmus ein Konstrukt mit vielen Facetten ist, die im NPI über Subskalen operationalisiert sind. Die beim NPI-10 vorgenommene Kürzung von 40 auf 10 Items stellt bereits eine hochgradige Verdichtung der Messung dar.

Zielgruppe der Testpersonen

Die aufgenommenen Tests sollen prinzipiell möglichst breit einsetzbar sein und nicht ausschließlich der Diagnostik spezifischer Gruppen dienen. Als Zielgruppe der Testpersonen wurde die erwachsene Bevölkerung festgelegt. Die Altersbegrenzung wurde, wie in zahlreichen Standardverfahren der Persönlichkeitsdiagnostik üblich (z. B. dem NEO-PI-R oder dem Freiburger Persönlichkeitsinventar FPI; Fahrenberg, Hampel & Selg, 2010), auf 16 Jahre angesetzt. Ausgeschlossen wurden Verfahren für das Säuglings-, Kleinkind-, Kindergarten-, Vorschul- und Schulalter sowie Verfahren, die ausschließlich für die Diagnostik bestimmter Gruppen entwickelt wurden.

Sprache

Die Zielgruppe dieses Buches sind deutschsprachige Leser. Daher wurde als Einschlusskriterium die deutsche Sprache festgelegt. Die Tests bzw. deren Items mussten entweder im deutschsprachigen Raum entwickelt oder nach einem gängigen Prozedere übersetzt worden sein. Sie mussten also in einer deutschen Fassung vorliegen. Die Qualität der Übersetzung war kein Selektionskriterium.

Verfügbarkeit

Die Tests müssen potenziellen Nutzern prinzipiell zugänglich sein. Das heißt, dass Instruktion und Items vollständig in digitaler oder kopierfähiger Form verfügbar sein müssen. Es wurde darauf geachtet, dass diese in Publikationen abgedruckt, im Internet veröffentlicht oder vom Testautor auf Anfrage erhältlich sind.

Psychometrische Mindeststandards

Als psychometrischer Mindeststandard wurde festgelegt, dass der Test nach 1989 an mindestens einer deutschen Stichprobe psychometrisch evaluiert worden sein muss. Zu jedem Test sollten Angaben zu den Hauptgütekriterien Objektivität, Reliabilität und Validität (vgl. Lienert & Raatz, 1998) vorliegen. Im Hinblick auf die Validität wurde den Autoren der Beiträge nahe gelegt, mindestens Angaben zur faktoriellen Validität bzw. internalen Struktur und zur konvergenten Validität zu machen. Dies war allerdings nur eine Empfehlung und kein Ausschlusskriterium.

Auf eine Festlegung bestimmter numerischer Mindestwerte (Grenzwerte) für die Kennwerte der psychometrischen Güte wurde verzichtet (vgl. Kersting, 2006). Die Höhe psychometrischer Kennwerte wird nicht vom Test determiniert, sondern resultiert aus der spezifischen Kombination eines Tests und der Stichprobe bzw. dem Kontext, in dem dieser eingesetzt wird (Fischer, 1968). Entscheidend sind hierbei die Streuungen der Testwerte. Zum Beispiel sind die Streuungen bei einem kognitiven Leistungstest in einer Stichprobe Studierender und einer bildungsrepräsentativen Bevölkerungsstichprobe stark unterschiedlich. Dies kann zu deutlichen Unterschieden in der Höhe von Reliabilitäts- und Validitätskoeffizienten

führen. Hinzu kommt, dass Reliabilitäts- und Validitätskoeffizienten Einzelschätzungen darstellen und daher messfehlerbedingten Schwankungen unterworfen sind (für eine detaillierte Diskussion von Nachteilen fester numerischer Ausprägungen von Testgütekriterien siehe Kersting, 2006).

Als Mindeststandard wurde daher definiert, dass die aufgenommenen Tests in mindestens einer empirischen Untersuchung psychometrisch überprüft wurden. Damit wurden zahlreiche ad-hoc-konstruierte Tests und Übersetzungen fremdsprachlicher Verfahren ausgeschlossen, die bisher teststatistisch nicht evaluiert wurden. Der Grad der psychometrischen Bewährtheit eines Tests war aus den oben diskutierten Gründen kein Ein- bzw. Ausschlusskriterium. Dies impliziert, dass der Leser sich selbst ein Urteil über die psychometrische Güte eines Tests bilden muss. In jedem Beitrag des Buches werden alle relevanten Angaben über einen Test systematisch zusammengestellt, um die notwendigen Voraussetzungen für eine solche Beurteilung zu schaffen. Detaillierte Hinweise zu den Testgütekriterien und deren Beurteilung finden sich bspw. bei Bühner (2011), Evers (2001) und Lienert und Raatz (1998).

Literatur

Barkmann, C., Schulte-Markwort, M. & Brähler, E. (2011). *Klinisch-psychiatrische Ratingskalen für das Kindes- und Jugendalter. Diagnostik für Klinik und Praxis: Vol. 6.* Göttingen: Hogrefe.

Binet, A. (1905). *L'Annee Psychologique, 12,* 191–244.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI) nach Costa und McCrae.* Göttingen: Hogrefe.

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion.* München: Pearson Studium.

Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214–227.

Costa, P. T. & McCrae, R. R. (1985). *The NEO personality inventory: Manual, form S and form R.* Odessa: Psychological Assessment Resources.

Coste, J., Guillemin, F., Pouchot, J. & Fermanian, J. (1997). Methodological approaches to shortening composite measurement scales. *Journal of Clinical Epidemiology, 50,* 247–252.

Credé, M., Harms, P., Niehorster, S. & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology, 102,* 874.

Doll, E. A. (1917). A brief Binet-Simon scale. *Psychological Clinic, 11,* 197–211.

Edwards, P., Roberts, I., Sandercock, P. & Frost, C. (2004). Follow-up by mail in clinical trials: does questionnaire length matter? *Controlled Clinical Trials, 25,* 31–52.

Evers, A. (2001). Improving test quality in the Netherlands: Results of 18 years of test ratings. *International Journal of Testing, 1,* 137–153.

Fahrenberg, J., Hampel, R. & Selg, H. (2010). *FPI-R: Freiburger Persönlichkeitsinventar; Manual.* Göttingen: Hogrefe.

Fischer, G. H. (1968). *Einführung in die Theorie psychologischer Tests: Grundlagen und Anwendungen.* Bern: Huber.

Gosling, S. D., Rentfrow, P. J. & Swann, W. B. (2003). A very brief measure of the Big Five personality domains. *Journal of Research in Personality, 37,* 504–528.

Kemper, C. J., Beierlein, C., Bensch, D., Kovaleva, A. & Rammstedt, B. (2012). *Eine Kurzskala zur Erfassung des Gamma-Faktors sozial erwünschten Antwortverhaltens: Die Kurzskala Soziale Erwünschtheit-Gamma (KSE-G)* (GESIS Working Papers 2012|25). Köln: GESIS.

Kemper, C. J., Lutz, J. & Neuser, J. (2011). Konstruktion und Validierung einer Kurzform der Skala Angst vor negativer Bewertung (SANB-5). *Klinische Diagnostik und Evaluation, 4,* 343–360.

Kersting, M. (2006). Zur Beurteilung der Qualität von Tests: Resümee und Neubeginn. *Psychologische Rundschau, 57,* 243–253.

Kruyen, P. M., Emons, W. H. & Sijtsma, K. (2013). On the shortcomings of shortened tests: A literature review. *International Journal of Testing, 13,* 223–248.

Levy, P. (1968). Short form tests: a methodological review. *Psychological Bulletin, 6,* 410–416.

Lienert, G. (1969). *Testaufbau und Testanalyse.* Weinheim: Beltz.

Lienert, G. & Raatz, U. (1998). *Testaufbau und Testanalyse.* Weinheim: Beltz.

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G. & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: Operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment, 17,* 81–102.

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66,* 574–583.

Nunnally, J. (1978). *Psychometric theory.* New York: McGraw-Hill.

Ostendorf, F. & Angleitner, A. (2004). *NEO-Persönlichkeitsinventar nach Costa und McCrae, Revidierte Fassung (NEO-PI-R).* Göttingen: Hogrefe.

Paunonen, S. V. & Jackson, D. N. (1985). The validity of formal and informal personality assessment. *Journal of Research in Personality, 19,* 331–342.

Pervin, L. A., Cervone, D. & John, O. P. (2005). *Persönlichkeitstheorien.* München: Ernst Rheinhardt.

Rammstedt, B. & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality, 41,* 203–212.

Rammstedt, B., Kemper, C. J. & Schupp, J. (in Druck). Leitartikel: Standardisierte Kurzskalen zur Erfassung psychologischer Merkmale in Umfragen. *Methoden – Daten – Analysen.*

Roberts, B. W., Kuncel, N. R., Shiner, R., Caspi, A. & Goldberg, L. R. (2007). The Power of Personality: The comparative validity of personality traits, socioeconomic status, and cognitive ability for predicting important life outcomes. *Perspectives on Psychological Science, 2* (4), 313–345.

Robins, R. W., Hendin, H. M. & Trzesniewski, K. H. (2001). Measuring global self-esteem: Construct validation of a single-item measure and the Rosenberg Self-Esteem Scale. *Personality and Social Psychology Bulletin, 27,* 151–161.

Rost, J. (2004). Zählen oder Messen. Methoden der Sozialforschung. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 44,* 75–92.

Smith, G. T., McCarthy, D. M. & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment, 12,* 102–111.

Stanton, J. M., Sinar, E. F., Balzer, W. K. & Smith, P. C. (2002). Issues and strategies for reducing the length of self-report scales. *Personnel Psychology, 55,* 167–194.

Stieglitz, R. D. (2007). *Diagnostik und Klassifikation in der Psychiatrie.* Stuttgart: Kohlhammer.

Thalmayer, A. G., Saucier, G. & Eigenhuis, A. (2011). Comparative validity of brief to medium-length Big Five and Big Six Personality Questionnaires. *Psychological Assessment, 23,* 995–1009.

Tupes, E. C. & Christal, R. C. (1961). *Recurrent personality factors based on trait ratings (Tech. Rep. No. ASD-TR-61-97).* Lackland Air Force Base, Texas: U. S. Air Force.

Wood, D., Nye, C. & Saucier, G. (2010). Identification and measurement of a more comprehensive set of person-descriptive markers in the English lexicon. *Journal of Research in Personality, 44,* 258–272.